Analyzing the Facebook friendship graph

S. Catanese\textsuperscript{1}, P. De Meo\textsuperscript{1,2}, E. Ferrara\textsuperscript{3}, G. Fiumara\textsuperscript{1} and A. Provetti\textsuperscript{1,4}

\textsuperscript{1}Dept. of Physics, Informatics Section, University of Messina
\textsuperscript{2}Dept. of Computer Sciences, Vrije Universiteit Amsterdam
\textsuperscript{3}Dept. of Mathematics, University of Messina
\textsuperscript{4}Oxford-Man Institute, University of Oxford

Int’l Conf. on Web Intelligence, Mining and Semantics
May 26th 2011, Sogndal
Outline

1 Motivation
   • Main objective
   • The Basic Problem
   • Classic Work

2 Our Results/Contribution
   • Data Extraction and Cleaning
   • Data Analysis
   • Main Results

3 Future Issues
Outline

1. Motivation
   - Main objective
   - The Basic Problem
   - Classic Work

2. Our Results/Contribution
   - Data Extraction and Cleaning
   - Data Analysis
   - Main Results

3. Future Issues
Main objective

- Extract a (partial) graph of friendship relations from Facebook
  - starting from the friendlist of a real user
  - accessing **only** publicly accessible data of Facebook users

- using:
  - a wrapper (for extraction, cleaning and normalization of data)
  - a tool for graph visualization and analysis

- developed by some of us
Main objective

- Extract a (partial) graph of friendship relations from Facebook
  - starting from the friendlist of a real user
  - accessing **only** publicly accessible data of Facebook users
- using:
  - a wrapper (for extraction, cleaning and normalization of data)
  - a tool for graph visualization and analysis
- developed by some of us
Main objective

- Extract a (partial) graph of friendship relations from Facebook
  - starting from the friendlist of a real user
  - accessing **only** publicly accessible data of Facebook users
- using:
  - a wrapper (for extraction, cleaning and normalization of data)
  - a tool for graph visualization and analysis
- developed by some of us
Outline

1 Motivation
   - Main objective
   - The Basic Problem
   - Classic Work

2 Our Results/Contribution
   - Data Extraction and Cleaning
   - Data Analysis
   - Main Results

3 Future Issues
Social Networks
A Taxonomy

Social Networks (SN)
Described with graphs representing users and relationships among them

- Organizational Networks
- Collaboration Networks
- Communication Networks
- Friendship Networks

Online Social Networks (OSNs) [1]:
- Social Communities: Facebook, MySpace, etc.
- Social Bookmarking: Digg, Delicious, etc.
- Content Sharing: YouTube, Flickr, etc.
Social Networks
A Taxonomy

Social Networks (SN)
Described with graphs representing users and relationships among them

- Organizational Networks
- Collaboration Networks
- Communication Networks
- Friendship Networks
- Online Social Networks (OSNs) [1]:
  - Social Communities: Facebook, MySpace, etc.
  - Social Bookmarking: Digg, Delicious, etc.
  - Content Sharing: YouTube, Flickr, etc.
Social Networks

Examples

**Figure:** Organizational Network

**Figure:** Friendship Network
Mining Online Social Networks

Motivation

- Is the **distribution** of friendship computable?
- Calculating **graph properties** of OSNs
- Exploiting **new algorithms** in following tasks:
  - Walking through a large graph (e.g. BFS, MHRW, etc.)
  - Data compression (matrix decomposition, quadtrees, etc.)
  - Efficient visualization of large graphs
  - Clustering data (Fruchterman-Reingold, Harel-Koren, etc.)
  - Optimize efficiency in metrics evaluation (e.g. All-Pairs Shortest-Paths related: BC, CC, diameter, etc.)
- Studying the **scalability** of the problem
- Investigating similarities between OSNs and real-life SNs
Mining Online Social Networks

Motivation

- Is the **distribution** of friendship computable?
- Calculating **graph properties** of OSNs
  - Exploiting **new algorithms** in following tasks:
    - Walking through a large graph (e.g. BFS, MHRW, etc.)
    - Data compression (matrix decomposition, quadtrees, etc.)
    - Efficient visualization of large graphs
    - Clustering data (Fruchterman-Reingold, Harel-Koren, etc.)
    - Optimize efficiency in metrics evaluation (e.g. All-Pairs Shortest-Paths related: BC, CC, diameter, etc.)
- Studying the **scalability** of the problem
- Investigating similarities between OSNs and real-life SNs
Mining Online Social Networks

Motivation

- **Is the distribution** of friendship computable?
- Calculating **graph properties** of OSNs
- Exploiting **new algorithms** in following tasks:
  - Walking through a large graph (e.g. BFS, MHRW, etc.)
  - Data compression (matrix decomposition, quadtrees, etc.)
  - Efficient visualization of large graphs
  - Clustering data (Fruchterman-Reingold, Harel-Koren, etc.)
  - Optimize efficiency in metrics evaluation (e.g. All-Pairs Shortest-Paths related: BC, CC, diameter, etc.)
- Studying the **scalability** of the problem
- Investigating similarities between OSNs and real-life SNs
Mining Online Social Networks

Motivation

- Is the **distribution** of friendship computable?
- Calculating **graph properties** of OSNs
- Exploiting **new algorithms** in following tasks:
  - Walking through a large graph (e.g. BFS, MHRW, etc.)
  - Data compression (matrix decomposition, quadtrees, etc.)
  - Efficient visualization of large graphs
  - Clustering data (Fruchterman-Reingold, Harel-Koren, etc.)
  - Optimize efficiency in metrics evaluation (e.g. All-Pairs Shortest-Paths related: BC, CC, diameter, etc.)
- Studying the **scalability** of the problem
- Investigating similarities between OSNs and real-life SNs
Mining Online Social Networks

Motivation

- Is the **distribution** of friendship computable?
- Calculating **graph properties** of OSNs
- Exploiting **new algorithms** in following tasks:
  - Walking through a large graph (e.g. BFS, MHRW, etc.)
  - Data compression (matrix decomposition, quadtrees, etc.)
  - Efficient visualization of large graphs
  - Clustering data (Fruchterman-Reingold, Harel-Koren, etc.)
  - Optimize efficiency in metrics evaluation (e.g. All-Pairs Shortest-Paths related: BC, CC, diameter, etc.)
- Studying the **scalability** of the problem
- Investigating similarities between OSNs and real-life SNs
Mining Online Social Networks
Pros and Cons

Pros:
- Large-scale studies of phenomena and behaviors impossible before
- Relations among users are clearly defined
- Data can be automatically acquired
- Huge amount of information can be mined
- Several levels of granularity can be established

Cons:
- Large-scale mining issues
- Computational and algorithmic challenges
- Online friendship $\neq$ Real-life friendship
- Bias of data depends on visiting algorithm [2]
Mining Online Social Networks
Pros and Cons

**Pros:**
- Large-scale studies of phenomena and behaviors impossible before
- Relations among users are clearly defined
- Data can be automatically acquired
- Huge amount of information can be mined
- Several levels of granularity can be established

**Cons:**
- Large-scale mining issues
- Computational and algorithmic challenges
- Online friendship ≠ Real-life friendship
- Bias of data depends on visiting algorithm [2]
Outline

1 Motivation
   - Main objective
   - The Basic Problem
   - Classic Work

2 Our Results/Contribution
   - Data Extraction and Cleaning
   - Data Analysis
   - Main Results

3 Future Issues
Classic Work on (online or offline) SNs

- Milgram, Travers [3]: the **Small World** problem (1969-70)
- Zachary [4]: ’mining’ and **modeling** real-life SNs (1980)
- Kleinberg [5]: the small world problem from an **algorithmic** perspective (2000)
- Golbeck et al. [6]: social networks vs OSNs (2005)
- Barabasi [7], Leskovec [8], Shneiderman [9], etc.: all focusing on OSNs and their analysis (nowadays)
  - Online Social Network Analysis and Tools
  - Large-scale data mining from OSNs
  - Visualization of large graphs
  - Bias of data acquired from OSNs
  - Dynamics and evolution of OSNs
Classic Work on (online or offline) SNs

- Milgram, Travers [3]: the **Small World** problem (1969-70)
- Zachary [4]: ’mining’ and **modeling** real-life SNs (1980)
- Kleinberg [5]: the small world problem from an **algorithmic perspective** (2000)
- Golbeck et al. [6]: social networks vs OSNs (2005)
- Barabasi [7], Leskovec [8], Shneiderman [9], etc.: all focusing on OSNs and their analysis (nowadays)
  - Online Social Network Analysis and Tools
  - Large-scale data mining from OSNs
  - Visualization of large graphs
  - Bias of data acquired from OSNs
  - Dynamics and evolution of OSNs
Classic Work on (online or offline) SNs

- Milgram, Travers [3]: the **Small World** problem (1969-70)
- Zachary [4]: ’mining’ and **modeling** real-life SNs (1980)
- Kleinberg [5]: the small world problem from an **algorithmic perspective** (2000)
- Golbeck et al. [6]: social networks vs OSNs (2005)
- Barabasi [7], Leskovec [8], Shneiderman [9], etc.: all focusing on OSNs and their analysis (nowadays)
  - Online Social Network Analysis and Tools
  - Large-scale data mining from OSNs
  - Visualization of large graphs
  - Bias of data acquired from OSNs
  - Dynamics and evolution of OSNs
Classic Work on (online or offline) SNs

- Milgram, Travers [3]: the **Small World** problem (1969-70)
- Zachary [4]: ’mining’ and **modeling** real-life SNs (1980)
- Kleinberg [5]: the small world problem from an **algorithmic perspective** (2000)
- Golbeck et al. [6]: social networks vs OSNs (2005)
- Barabasi [7], Leskovec [8], Shneiderman [9], etc.: all focusing on OSNs and their analysis (nowadays)
  - Online Social Network Analysis and Tools
  - Large-scale data mining from OSNs
  - Visualization of large graphs
  - Bias of data acquired from OSNs
  - Dynamics and evolution of OSNs
Classic Work on (online or offline) SNs

- Milgram, Travers [3]: the **Small World** problem (1969-70)
- Zachary [4]: ’mining’ and **modeling** real-life SNs (1980)
- Kleinberg [5]: the small world problem from an **algorithmic perspective** (2000)
- Golbeck et al. [6]: social networks vs OSNs (2005)
- Barabasi [7], Leskovec [8], Shneiderman [9], etc.: all focusing on OSNs and their analysis (nowadays)
  - Online Social Network Analysis and Tools
  - Large-scale data mining from OSNs
  - Visualization of large graphs
  - Bias of data acquired from OSNs
  - Dynamics and evolution of OSNs
Outline

1 Motivation
   - Main objective
   - The Basic Problem
   - Classic Work

2 Our Results/Contribution
   - Data Extraction and Cleaning
   - Data Analysis
   - Main Results

3 Future Issues

Catanese, De Meo, Ferrara, Fiumara & Provetti (2011) Analyzing the Facebook friendship graph
**Mining the Facebook graph**

**Visiting Algorithm**

**BFS approach**: starting from a single seed (a FB profile), visiting friend-lists of nodes in order of discovering.

**Pros:**
- Optimal solution for unw. und. graphs
- Implementation is easy and intuitive

**Cons:**
- Introduces bias in incomplete visits

**Challenges:**
- FB anti-data mining policies

**Figure**: Breadth-first search (3rd sub-level)

1. Seed
2-4 Friends
5-8 Friends of friends
9-12 Friends of fr. of fr.
Mining the Facebook graph

Design of the Mining Agent

Figure: State Diagram of the Data Mining Process
Mining the Facebook graph

Architecture
- Java application
- Firefox browser embedded
- XPCOM/XULRunner interface
- Web pages spider
- Wrapper

fig10.png
Mining the Facebook graph

How the Agent Works

Agent Initialization:
- FB authentication → Seed friend-list page
- Selection an example friend → XPath extraction
- Wrapper generation and adaptation
- Wrapper execution → Generation of the queue

Agent Execution:
- Load FIFO queue
- For all the user profiles in the queue:
  ▶ Visit friend-list page of the current user
    ★ Extract friends (nodes) and save friendships (edges)
    ★ Insert unvisited profiles in the queue
  ▶ Visit ’next pages’ of the friend-list
  ▶ Cycle the process
Mining the Facebook graph

Handling Data

Possible representations of visited nodes and edges:

- *Adjacency list*
- Adjacency matrix

Possible representation of BFS visit for unvisited nodes:

- FIFO queue
Mining the Facebook graph

Cleaning Data
- removing duplicate nodes exploiting hash tables
- relinking edges
- deleting parallel edges

**Data cleaning**: $O(n)$ time (optimal)

**Structured Format**: Clean data is saved under the XML structure GraphML

fig9.png

fig6.png
Mining the Facebook graph

Agent Running

fig7.png
Outline

1 Motivation
   - Main objective
   - The Basic Problem
   - Classic Work

2 Our Results/Contribution
   - Data Extraction and Cleaning
   - Data Analysis
   - Main Results

3 Future Issues
Network Analysis Metrics

Types of Networks

Classifications of several types of networks exist. They affect metrics and maps generated in order to reflect their interpretation.

- **Networks member’s point of view**
  - Egocentric
  - Partial
  - Full

- **Networks entity’s point of view**
  - Unimodal
  - Multimodal
  - Bimodal
  - Affiliation

- Multiplex networks
Classifications of several types of networks exist. They affect metrics and maps generated in order to reflect their interpretation.

- **Networks member’s point of view**
  - Egocentric
  - Partial
  - Full

- **Networks entity’s point of view**
  - Unimodal
  - Multimodal
  - Bimodal
  - Affiliation

- **Multiplex networks**
Classifications of several types of networks exist. They affect metrics and maps generated in order to reflect their interpretation.

- **Networks member’s point of view**
  - Egocentric
  - Partial
  - Full

- **Networks entity’s point of view**
  - Unimodal
  - Multimodal
  - Bimodal
  - Affiliation

- **Multiplex networks**
Network Analysis Metrics

Facebook Friendship Network

Facebook characteristics:

- **Egocentric** networks: the term ego denotes a person connected to everyone (alter) in the network
- **Unweighted, undirected** network:

  - 1.0 degree
  - 1.5 degree
  - 2.0 degree

![fig12.png](attachment:fig12.png)

Shows a natural effect of clustering around different areas of a person's life: friends, classmates, workmates, family, etc.

Catanese, De Meo, Ferrara, Fiumara & Provetti (2013)

Analyzing the Facebook friendship graph
Network Analysis Metrics

Facebook Friendship Network

*Facebook characteristics:*

- **Egocentric** networks: the term ego denotes a person connected to everyone (alter) in the network
- **Unweighted, undirected** network:

  - 1.0 degree
  - 1.5 degree
  - 2.0 degree

![fig12.png](https://example.com/fig12.png)
### Network Analysis Metrics

#### Facebook Friendship Network

**Facebook characteristics:**
- **Egocentric networks**: the term ego denotes a person connected to everyone (alter) in the network
- **Unweighted, undirected network**:

<table>
<thead>
<tr>
<th>Degree</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td></td>
</tr>
<tr>
<td>1.5</td>
<td></td>
</tr>
<tr>
<td>2.0</td>
<td></td>
</tr>
</tbody>
</table>

- fig12.png

Catanese, De Meo, Ferrara, Fiumara & Provetti (2011)
Network Analysis Metrics

Measures

Network metrics

Allow analysts to systematically dissect the social world, creating a basis on which to compare networks, track changes in a network over time and determine the relative position of individuals and clusters within the network.

Research focuses on:
- Structure of the whole graph;
- Large sub-graphs;
- Identifying individual nodes of particular interest;
- Analyze the whole graph aggregated over its entire lifetime;
- To slice the network into units of time to explore the progression of the development of the network.

A starting point: list from Perer and Shneiderman
Network Analysis Metrics

Network metrics

Allow analysts to systematically dissect the social world, creating a basis on which to compare networks, track changes in a network over time and determine the relative position of individuals and clusters within the network.

- Research focuses on:
  - Structure of the whole graph;
  - Large sub-graphs;
  - Identifying individual nodes of particular interest;
  - Analyze the whole graph aggregated over its entire lifetime;
  - To slice the network into units of time to explore the progression of the development of the network.

- A starting point: list from Perer and Shneiderman
Network Analysis Metrics
Measures - Perer and Shneiderman List

- **Overall network metrics**: number of nodes, number of edges, density, diameter ecc;
- **Node rankings**: degree, betweenness and closeness centrality;
- **Edge rankings**: weight, betweenness centrality;
- **Node rankings in pairs**: degree vs. betweenness, plotted on a scatter gram;
- **Edge rankings in pairs**;
- **Cohesive subgroups**: finding communities;
- **Multiplexity**: analyzing comparisons among different edge types.
Visualizing Social Networks

Constructing visual images of social networks provides insights about the structure of a network, so as representing a visual support for explaining network phenomena [10].

- **Graph drawing issues:**
  - As network complexity increases, its illegibility increases as well;
  - Interactive operations on nodes, such as filtering or manual placement, are needed.
Readability Metrics (RMs)

RMs measure how much understandable is the graph drawing (such as the number of edge crossings or occluded nodes in the drawing) [11].

- Each algorithm attempts to find an optimal layout of the graph, often according to a set of readability metrics;
  - A simple interim set of guidelines might aspire to the four principles of *NetViz Nirvana* [12]:
    - Every vertex is visible;
    - Every vertex’s degree is countable;
    - Every edge can be followed from source to destination;
    - Clusters and outliers are identifiable.

- **Approach**: layout and filtering techniques.
Analyzing Social Networks
Better-quality Network Visualization

Readability Metrics (RMs)
RMs measure how much understandable is the graph drawing (such as the number of edge crossings or occluded nodes in the drawing) [11].

- Each algorithm attempts to find an optimal layout of the graph, often according to a set of readability metrics;
- A simple interim set of guidelines might aspire to the four principles of NetViz Nirvana [12]:
  - Every vertex is visible;
  - Every vertex’s degree is countable;
  - Every edge can be followed from source to destination;
  - Clusters and outliers are identifiable.

Approach: layout and filtering techniques.
Readability Metrics (RMs)

RMs measure how much understandable is the graph drawing (such as the number of edge crossings or occluded nodes in the drawing) [11].

- Each algorithm attempts to find an optimal layout of the graph, often according to a set of readability metrics;
- A simple interim set of guidelines might aspire to the four principles of NetViz Nirvana [12]:
  - Every vertex is visible;
  - Every vertex’s degree is countable;
  - Every edge can be followed from source to destination;
  - Clusters and outliers are identifiable.
- **Approach:** layout and filtering techniques.
SNA Tools
Some Powerful Tools and Libraries Adopted

- **GUESS** focuses on improving the interactive exploration of graphs.
- **NodeXL** developed as an add-in to the Microsoft Excel 2007 spreadsheet software, provides tools for network overview, discovery and exploration.
- **LogAnalysis** helps forensic analysts in visual statistical analysis of mobile phone traffic networks.
- **Jung** and **Prefuse** provide Java APIs implementing algorithms and methods for building applications for graphical visualization and SNA for graphs.
- A list of other SNA tools for extract, analyze and display social media networks can be found on International Network for Social Network Analysis (INSNA) site \(^1\).

\(^1\)http://www.insna.org/software/index.html
Outline

1 Motivation
   - Main objective
   - The Basic Problem
   - Classic Work

2 Our Results/Contribution
   - Data Extraction and Cleaning
   - Data Analysis
   - Main Results

3 Future Issues
### Facebook Network Analysis
**NodeXL - Overall Metrics**

<table>
<thead>
<tr>
<th>Metric</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Graph Type</td>
<td>Undirected</td>
</tr>
<tr>
<td>Vertices</td>
<td>547,302</td>
</tr>
<tr>
<td>Unique Edges</td>
<td>836,468</td>
</tr>
<tr>
<td>Edges With Duplicates</td>
<td>0</td>
</tr>
<tr>
<td>Total Edges</td>
<td>836,468</td>
</tr>
<tr>
<td>Self-Loops</td>
<td>0</td>
</tr>
<tr>
<td>Connected Components</td>
<td>2</td>
</tr>
<tr>
<td>Single-Vertex Connected Components</td>
<td>0</td>
</tr>
<tr>
<td>Maximum Vertices in a Connected Component</td>
<td>546,733</td>
</tr>
<tr>
<td>Maximum Edges in a Connected Component</td>
<td>835.9</td>
</tr>
<tr>
<td>Maximum Geodesic Distance (Diameter)</td>
<td>10</td>
</tr>
<tr>
<td>Average Geodesic Distance</td>
<td>5.00</td>
</tr>
</tbody>
</table>

**Table:** Overall Network Metrics
Facebook Network Analysis
NodeXL - Miscellaneous Metrics

<table>
<thead>
<tr>
<th>Metric</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Average</th>
<th>Median</th>
</tr>
</thead>
<tbody>
<tr>
<td>Degree</td>
<td>1</td>
<td>4,958</td>
<td>3.057</td>
<td>1.000</td>
</tr>
<tr>
<td>PageRank</td>
<td>0.269</td>
<td>2,120.268</td>
<td>1.000</td>
<td>0.491</td>
</tr>
<tr>
<td>Clustering Coefficient</td>
<td>0.000</td>
<td>1.000</td>
<td>0.053</td>
<td>0.000</td>
</tr>
<tr>
<td>Eigenvector Centrality</td>
<td>0.000</td>
<td>0.003</td>
<td>0.000</td>
<td>0.000</td>
</tr>
</tbody>
</table>

Table: Miscellaneous Metrics
Facebook Network Graph
LogAnalysis Force Directed Filtered View (25K Nodes Sub-graph)

fig7cat.png
Facebook Network Graph
LogAnalysis Force Directed Filtered View (2.0 degree)

fig8cat.png
Facebook Network Graph
LogAnalysis Force Directed Aggregate Filtered View (2.0 Degree)

fig9cat.png
Facebook Network Graph

NodeXL Visualization (25K Nodes Sub-graph)

fig10cat.png
Facebook Network Graph
NodeXL Filtered Visualization (25K Nodes Sub-graph)

fig11cat.png
Facebook Network Graph
NodeXL Filtered Visualization (25K Nodes Sub-graph)

fig12cat.png
Metrics Importance: Betweenness Centrality
Top 25 Nodes Ordered by BC (25K Nodes Sub-graph)

fig4.png
Future Issues

- Enrich the sample (currently 5 million nodes and 15 million edges)
- Refine features and metrics thanks to larger sample
- Study communities emerging from the overall graph
- Implement parallel techniques to speed-up metrics calculations
- Determine scaling (-up and -down) coefficients
- How visiting algorithms affect extracted data
- Dynamic (i.e., temporal) evolution of the graph
Thank you
For Further Reading I

R. Kumar
Online social networks: modeling and mining
*Proc. of the 2nd ACM International Conference on Web Search and Data Mining*, 2009

M. Kurant, A. Markopoulou, P. Thiran
On the bias of BFS

S. Milgram, J. Travers
An experimental study of the small world problem
*Sociometry, 32(4), 1969*

W. Zachary
A language for modeling and simulating social process
*PhD Thesis, 1980*
For Further Reading II

J. Kleinberg
The small-world phenomenon: an algorithm perspective
Proc. of the 32nd ACM symposium on Theory of computing, 2000

J. Golbeck et al.
Social networks applied
IEEE Intelligent Systems, 20(1), 2005

A.L. Barabasi et al.
Linked: the new science of networks
American Journal of Physics, 71(4), 2003

J. Leskovec
Dynamics of large networks
For Further Reading III

B. Shneiderman
Analyzing (social media) networks with NodeXL
Proc. of the 4th International Conference on Communities and Technologies, 2009

L.C. Freeman
Visualizing Social Networks
Journal of Social Structure, 2000

B. Shneiderman, C. Dunne
Improving Graph Drawing Readability by Incorporating Readability Metrics: A Software Tool for Network Analysts
University of Maryland, HCIL Tech Report HCIL-2009-13, May 2009

B. Shneiderman, A. Aris
Network Visualization with Semantic Substrates
For Further Reading IV